Nonlin. Processes Geophys., 15, 115–126, 2008
www.nonlin-processes-geophys.net/15/115/2008/

Nonlinear Processes in Geophysics

Estimation of soil types by non linear analysis of remote sensing data

C. Hahn and R. Gloaguen

Remote Sensing Group, Geology Institute, TU Bergakademie Freiberg, 09599 Freiberg, Germany

Received: 3 September 2007 – Revised: 6 December 2007 – Accepted: 19 December 2007 – Published: 15 February 2008

Abstract. The knowledge of soil type and soil texture is crucial for environmental monitoring purpose and risk assessment. Unfortunately, their mapping using classical techniques is time consuming and costly. We present here a way to estimate soil types based on limited field observations and remote sensing data. Due to the fact that the relation between the soil types and the considered attributes that were extracted from remote sensing data is expected to be non-linear, we apply Support Vector Machines (SVM) for soil type classification. Special attention is drawn to different training site distributions and the kind of input variables. We show that SVM based on carefully selected input variables proved to be an appropriate method for soil type estimation.

1 Introduction

Soils play an important role in the environment. They are natural habitats for flora and fauna, determine plant growths, store water, filter and/or transform infiltrating substances. Soil texture strongly influences, for example, the water holding capacity, stability, erodability, and permeability of soil. Soil type is important for agriculture, forestry and planning purposes as it reveals knowledge about soil horizons, information about ground water and back water influence and leaching processes. Unfortunately, soil type and soil texture mapping is a very time and cost consuming task. Based on the desire to reduce expensive field observations, the aim of this work is to estimate soil types using limited ground data and additional attributes. Due to the fact that the relation between the soil types and the additional attributes is expected to be non-linear, Support Vector Machines (SVM) were used. Our experience on the performance of linear analyses (e.g. logistic regression) supported the choice of a non-linear classifier. Unlike Bhattacharya and Solomatine (2006) who based soil texture classification on Cone Penetration Testing (CPT) data, and Pozdnukhov et al. (2002) who solely used coordinates we propose to handle attributes that can be derived from satellite images as input variables for soil type SVM classification. Thus, the following input variables were considered: landuse, altitude, aspect, slope, geology, distance to rivers and coordinates. The study area is located in the eastern part on the Erzgebirge, Germany. Different classifications were perfomed in order to address the following three questions: (I) How does training site distribution affect the performance of soil type classification using SVM. (II) Should the coordinates always be considered as input variables. (III) Does a sine and cosine representation of the aspect return better results than the degree representation of the aspect.

Correspondence to: C. Hahn
(claudiahahn@yahoo.de)

2 Main principles of Support Vector Machines (SVM)

This section gives a brief introduction on the basic principles of SVM classification. More detailed information about SVM and their mathematical background, can be found in Schölkopf and Smola (2002), Chen et al. (2005), Burges (1998), Shawe-Taylor and Cristianini (2000) and Vapnik (2000). SVM is a type of universal learning machine (Vapnik, 2000). In other words it is a learning algorithm used for pattern recognition and classification and was originally designed to solve binary classification problems. Therefore, the performance of SVM will be described using a binary classification problem. The linear, non-linear and the non-separable cases and at last the multi-class case are addressed.

2.1 Linear SVMs for the separable case

Consider a training data set

$$\{(\boldsymbol{x_i}, y_i) \mid \boldsymbol{x_i} \in X \subset R^n, y_i \in \{-1, 1\}\} \quad (1)$$

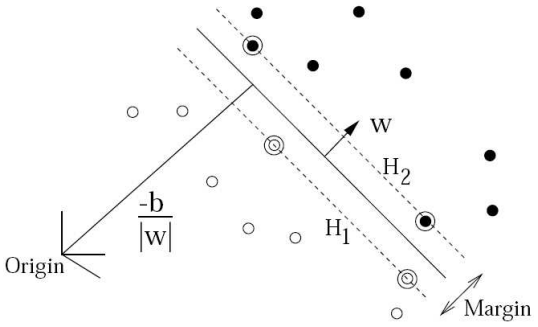


Fig. 1. Linear separating hyperplane for the separable case (Burges, 1998).

where $\boldsymbol{x_i}$ are input data points (called patterns) in a n-dimensional space (the input space X), that are labeled by y_i. The variable y_i defines the class membership of each point, and in this case y_i can either be 1 or -1. For example, $\boldsymbol{x_1}\in\mathrm{R}^3$ denotes one point in a three dimensional space, constructed for instance by the first three spectral bands of Landsat, hereafter referred to as attributes or input variables, and is labeled by y_1=1, which could code the landuse class forest. SVM solve the classification problem by constructing an optimal hyperplane that separates the data (Fig. 1). The class of hyperplanes considered and the corresponding decision functions look as follow (Chen et al., 2005): Class of hyperplanes

$$(\boldsymbol{w}\cdot\boldsymbol{x})+b=0 \qquad \boldsymbol{w}\in R^n, b\in R \tag{2}$$

Decision functions

$$f(x)=sgn((\boldsymbol{w}\cdot\boldsymbol{x})+b) \tag{3}$$

The vector $\boldsymbol{w}$ is perpendicular to the hyperplane and called a weight vector. b is called a treshold or bias and $|b|/||\boldsymbol{w}||$ is the perpendicular distance from the hyperplane to the origin (Burges, 1998), with $||\boldsymbol{w}||$ beeing the Euclidean norm of $\boldsymbol{w}$. (Fig. 1). b=0 would force the hyperplane to pass through the origin.

The points closest to the optimal hyperplane are situated on the parallel hyperplanes H_1 and H_2 and are named support vectors. In Fig. 1 they are circled.

To find the optimal hyperplane, training points that are closest to the hyperplane must be identified, and $\boldsymbol{w}$ and b must be choosen such that the margin is maximised. The margin, measured perpendicular to the hyperplane, is the sum of the distances of the closest point of each class to the hyperplane (Burges, 1998). Since the margin is proportional to $1/||\boldsymbol{w}||$ the optimisation problem can be expressed as follows:

$$\begin{aligned} &\text{minimise} && \tfrac{1}{2}||\boldsymbol{w}||^2 \\ &\text{subject to} && y_i\cdot((\boldsymbol{w}\cdot\boldsymbol{x_i})+b)\geq+1 \; i=1,...,m \end{aligned} \tag{4}$$

in other words

$$\begin{aligned} (\boldsymbol{w}\cdot\boldsymbol{x_i})+b\geq+1, \; for \; y_i=+1 \\ (\boldsymbol{w}\cdot\boldsymbol{x_i})+b\leq-1, \; for \; y_i=-1 \end{aligned}$$

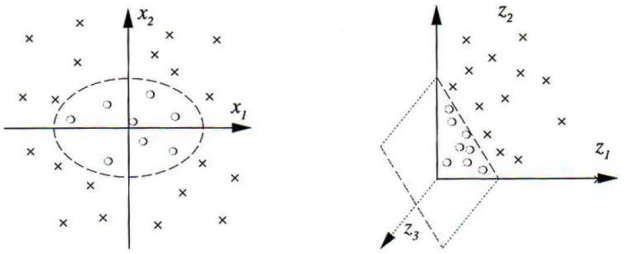


Fig. 2. Mapping of input variables into a higher dimensional space. Left: input space; Right: feature space (Schölkopf and Smola, 2002).

The optimal hyperplane is thus characterised by the largest margin between the two classes, and the ability to correctly separate the data. This optimisation problem can be solved through its Lagrangian dual, which leads to the following decision function, with α_i being dual variables (Chen et al., 2005):

$$f(x)=sgn(\sum_{i=1}^{m}y_i\alpha_i\cdot(\boldsymbol{x}\cdot\boldsymbol{x_i})+b). \tag{5}$$

It needs to be pointed out that only a subset of the training patterns, the support vectors, are relevant to construct the optimal hyperplane. The classification of a new data point $\boldsymbol{x}$ is based on a weighted comparison between this point and the support vectors. The dot product is used as a similarity measure.

2.2 Non-linear SVM for the separable case

Most of the time it is not possible to separate the data in the input space X using a linear function. Therefore, the idea behind SVM is to map the input data into a higher dimensional space (feature space H), where a hyperplane that separates the two classes can be constructed (Fig. 2). The example in Fig. 2 shows that the hyperplane in the feature space corresponds to a non-linear surface in the input space.

$$\begin{aligned} &\phi X\rightarrow H \\ &\boldsymbol{x}\rightarrow\phi(\boldsymbol{x}) \end{aligned} \tag{6}$$

Using a kernel k, it is possible to calculate the dot product in feature space without mapping all input data points into a higher dimensional space. Thus, using the kernel k as a similarity measure, the computational effort can be reduced. The decision function needs now to be changed to (Chen et al., 2005)

$$f(x)=sgn(\sum_{i=1}^{m}y_i\alpha_i\cdot k(\boldsymbol{x},\boldsymbol{x_i})+b) \tag{7}$$

Widely used kernels are the polynomial kernel or the Gaussian Radial Basis Function kernel (RBF:

$K(x_i, x_j)$=exp$(-||x_i - x_j||^2/2\sigma^2))$. Depending on the kernel the user has to define certain kernel parameters. In case of RBF, one needs to determine σ.

2.3 Non-separable case

In practice, a separating hyperplane often does not exist due to a large overlap of the classes. Thus, to define a hyperplane it is necessary to relax the constraints Eq. (4). Therefore, slack variables ξ_i were introduced (Cortes and Vapnik, 1995)

$$\xi_i \geq 0, \text{ where } i = 1, ..., m \tag{8}$$

Relaxing the constraints to:

$$y_i \cdot ((\boldsymbol{w} \cdot \boldsymbol{x_i}) + b) \geq 1 - \xi_i, i = 1, ..., m \tag{9}$$

The aim of this approach is to find a hyperplane that maximises the margin and keeps the misclassification of training examples small. "The trade-off between the margin and the misclassification error is controlled by a user-defined constant" (Pal and Mather, 2005). Applying the "Soft Margin Support Vector Classifiers" C-SVC the user needs to define a parameter C, with $C>0$, which controls the trade off.

$$\text{minimise } \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

subject to the constraints Eqs. (8) and (9)

A large C forces the creation of an accurate model, with very few misclassifications that may not generalise well (Foody and Mathur, 2004). Choosing a smaller C allows more misclassifications. C is called a regularisation parameter.

2.4 Multi-class case

Two main approaches have been developed for multi-class classification using SVM: The one-against-all (OAA) and one-against-one (OAO) method. OAA compares one class with all the others taken together. Having n classes, n hyperplanes are determined, n optimisation problems need to be solved and n classifiers are generated. The OAO approach performs a binary SVM on all possible pairs out of n classes. Each classifier is trained on two out of n classes. The number of classifiers therefore is n(n-1)/2. Applying these classifiers to a test data point leads to n(n-1)/2 class votes. The test data gets the class label from the class that received the most votes. Thus, in comparison to the OAA approach, more classifiers have to be generated when the OAO method is applied. On the other hand, the OAA approach suffers from problems caused by unbalanced training set sizes (Foody and Mathur, 2004), (Huang et al., 2002). Both methods, OAA and OAO, reduce the multiclass dataset to several binary problems that have to be solved.
Some new approaches were developed to solve the multiclass problem without reducing it to several binary cases (e.g. Weston and Watkins, 1998; Hsu and Lin, 2002; Foody and Mathur, 2004).

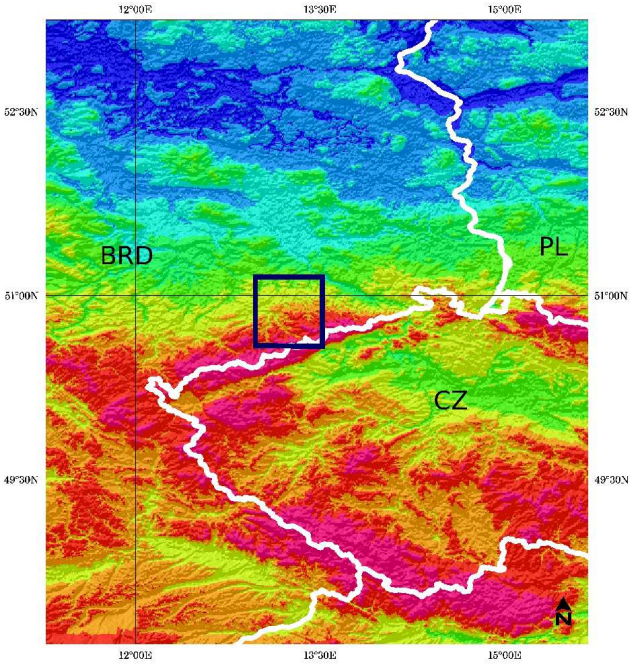


Fig. 3. Location of study area.

3 Study area and experimental design

3.1 Study area

To investigate the performance of SVM for soil type classification, with regards to training site distribution and certain input variables,a study area was chosen where an exhaustive data base was available. The area under investigation is situated in the eastern part of the Erzgebirge/Germany and is part of the catchment area of the river Weißeritz (Fig. 3).
This region has a size of about 28 km^2 with an altitude ranging from 544 m to about 800 m above sea level. The topographically undulating area is dominated by fields, grassland and coniferous forest. Urban areas are mainly located along roads and are characterised by small houses surrounded by gardens. Within the study region 12 soil types were distinguished using the German classification scheme. An accurate transformation into international terms was difficult. Thus, only the main soil types are named in the following and afterwards we refer to the different soil types using labels 1 to 12. Figure 4 shows the distribution of the 12 soil types within the study area. The dominating soil types are *Cambisols* (label:3) and *Planosol – Cambisols* (labels:4,5). *Planosols* (label:6), *Gleysols* (label:9), *Podzols* (label:2), *Fluvisols* (label:8), *Histosols* (label:10) and some transitional soil types (labels:7,12) also occur in this area, but occupy considerably less space. In regions covered with forest the underlying soil type is mainly *Cambisol*, whereas grassland and fields are mostly located on *Planosols*/*Cambisols*. *Gleysols*

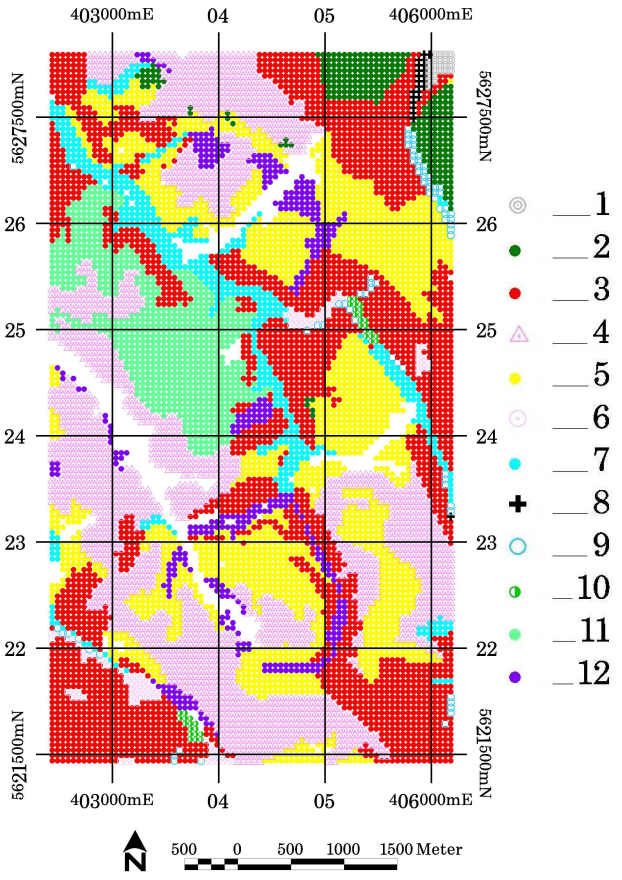


Fig. 4. Distribution of soil types within the study area (based on soil maps BK 5247 and BK 5248 provided by the Saxonian State Office of Environment and Geology): Soil classification based on "Bodenkundliche Kartieranleitung KA4": 1: BB-PP; 2: PPn; 3: BBn; 4: 2SS-BB+3BBn; 5: 3SS-BB+2BBn; 6: SSh; 7: 1GG-SS+2SS-GG+2GGn; 8: AB; 9: GG; 10: HH; 11: 2BB-SS+2SS-BB+1SSn; 12: 2ABn+3GG-AB.

and *Fluvisols* occur around rivers and streams. Geologically, gneiss, muscovite gneiss, albite and quarz phyllite and porphyry are dominant within the study area.

3.2 SVM soil type classification

Soil formation processes and the relation between site characteristics and soil types are very complex. Therefore, as expected and demonstrated later in this paper, linear classifiers perform poorly. We decided to apply SVM, as it is able to handle non-linear as well as non-separable cases.

3.2.1 SVM settings

Soil type classifications were performed using RBF kernel and the One-Against-One multi-class approach. The RBF kernel was chosen because this kernel is able to handle non-linear relations between class labels and attributes, and only one kernel parameter has to be defined (Hsu et al., 2007). Furthermore, Kanevski et al. (2007) state that the Gaussian Radial Basis Functions (RBF) are suitable for environmental applications. The decision to use the OAO approach was based on the experience that the OAO method yields better results than the OAA approach (e.g. Pal and Mather, 2005) and to avoid problems caused by unbalanced training set sizes (refer to Sect. 2.4). Furthermore, we applied the Soft-Margin Support Vector Classifier C-SVC. The kernel parameter σ and the regularisation parameter C were determined using 10-fold cross-validation (CV). In case several parameter combinations led to the smallest average error, the one leading to the smallest test error was chosen.

3.2.2 Input variables

Since soil type is closely related to soil formation, soil type classification was based on variables known to influence soil formation: altitude, aspect, slope, geology, distance to rivers and streams, and landuse. These variables can be derived from remote sensing data. In the present study the land use data were obtained from SVM classifications based on ASTER visible and shortwave infrared data. For some classifications coordinates were additionally used as input variables. The input of geographical coordinates were proved valuable for soil texture mapping (Pozdnukhov et al., 2002) and are assumed to enhance modelling of the spatial distribution of the soil types. Special attention was also drawn to the input variable aspect. Aspect is commonly represented in degree. Thus, very small and very large values represent a northern aspect. This would cause problems for classifiers using descriptive statistics, such as mean and standard deviation, to separate the data. However, it should be no problem for SVM, as they are not based on descriptive statistics and can handle non-linear cases. Sine and cosine of the angle are able to describe the aspect explicitly and are thus used here to investigate whether it holds true that SVM can handle the degree representation of the aspect. Combinations of input variables considered within this work are listed in Table 1.
All input variables were linearly scaled to range from -1 to 1. Regarding categorical attributes, like landuse and geology, each category was represented as a vector of -1 and 1. For example, three different landuse classes were represented as follows: forest (1,$-$1,$-$1), field ($-$1,1.$-$1), grassland ($-$1,$-$1,1).

3.2.3 Considered distributions of training data

For any supervised classification method an efficient training site selection is crucial. Typically, training site selection aims to provide an accurate description of each class in the sense of descriptive statistics. To achieve this, a rather large training data set is generally required. However, Foody and Mathur (2006) point out that the training data should provide the classifier with information neccessary to separate the classes, not to describe the classes. Hence, different

Table 1. Input variable combinations considered in this study.

Label	Input variables
A	altitude, slope, landuse, distance to rivers, geology, degree representation of the aspect
B	altitude, slope, landuse, distance to rivers, geology, sine and cosine representation of the aspect
C	altitude, slope, landuse, distance to rivers, geology, degree representation of the aspect, coordinates
C*	altitude, slope, landuse, distance to rivers, geology, sine and cosine representation of the aspect, coordinates

classifiers may require different training data, depending on how they separate the data. In contrast to MDC or MLC, SVM only use marginal data, the support vectors, to construct the hyperplane, which separates the data. Thus, the training data set should especially contain data points located close to the hyperplane. Based on this knowledge, Foody and Mathur (2006) present a new training data selection method for remote sensing data which would be easier and cheaper in comparison to conventional methods. They focus on mixed pixels located at class boundaries and tested their method on classifications of agricultural crops.

However, this approach cannot realistically be applied for soil type classification because, firstly, no real class boundaries exist, but rather transitions from soil type to soil type and, secondly, one cannot observe soil types surfaces in the field as soils are usually covered by vegetation. It is not possible to determine the location of class boundaries based on just the few ground data points that are available. Within this study, training data points were randomly or evenly selected. Soil type classifications were performed on four different training data sets (Fig. 5) to investigate the effect of training site distribution for SVM classification of soil types.

The first training data set is characterised by an almost uniform distribution of data points spread over the whole study area (Fig. 5A). Training set 2 is based on a random sampling of 1/14 of all data points (Fig. 5B). In training set 3 and 4 the data points are not spread over the whole study area. Training set 3 is arranged in a chequerboard like pattern, i.e. data points are randomly distributed in three distinct parts of the study area (Fig. 5C). Training set 4 is evenly distributed in the southern part of the study area (Fig. 5D). For all four training data sets the corresponding test data sets were large enough to enable good accuraccy assessment. However, the test data set belonging to training set 3 contains no samples of classes 6 and 10. Thus, it could not be tested if the classifier would be able to assign these two class memberships correctly.

Looking at Table 2 it is obvious that the training data sets are highly unbalanced. Such unbalanced data sets are unfavourable for SVM classification, and problematic for the CV approach as optimization parameters depend on the average CV error. Having an unbalanced data set with some classes comprising only one to ten and other more than 1000 data points (Table 2, training set 4), only the large classes determine the average error and thus parameter selection. However, the unbalanced distribution of training data points between the soil types is a realistic representation of the overall class distribution in the whole study area. Real field measurement campaigns would lead to similar unbalanced samples. Thus, training sets presented above were used for classification without modification. Training set 2 is the most realistic training set, in case the area under investigation is easily accessible. Training set 3 and 4 represent data sets acquired in regions which are only partly accessible.

Table 2. Number of training and test data in the four training sets. This table describes the unbalanced repartition of the training sets.

	Training set 1		Training set 2		Training set 3		Training set 4	
soil type	train	test	train	test	train	test	train	test
1	8	16	3	21	0	24	0	24
2	87	190	20	257	2	249	5	272
3	943	1887	201	2629	222	1070	1687	1143
4	794	1594	152	2236	155	1360	1752	636
5	710	1444	156	1998	138	1088	1313	841
6	36	70	17	89	15	0	78	28
7	184	358	45	497	50	79	228	314
8	8	15	2	21	1	22	1	22
9	23	50	3	70	6	41	23	50
10	13	23	5	31	3	0	21	15
11	265	525	43	747	28	623	390	400
12	147	263	31	379	30	229	253	157
total	3218	6435	678	8975	650	4785	5751	3902

3.3 Minimum Distance Classification (MDC) and Maximum Likelihood Classification (MLC)

To demonstrate the need for a non-linear classifier for soil type classification, MDC and MLC were performed as well. MLC, which in general yields better results than MDC, was only performed on numerical input variables. For categorical input variables the presupposition of normal distribution cannot be fulfilled. Leaving out such categorical data reduced classification performance for SVM significantly.

MLC and MDC were applied to training set 2, the most realistic training data set, in case the study area is easily accessible. The input variable combinations C and C* (Table 1) were considered for MDC to investigate whether a sine and cosine representation yields indeed better results than a degree representation of the aspect. All input variables were, as for SVM, scaled to range between -1 and 1. Categorical input variables were also treated the same way as for SVM

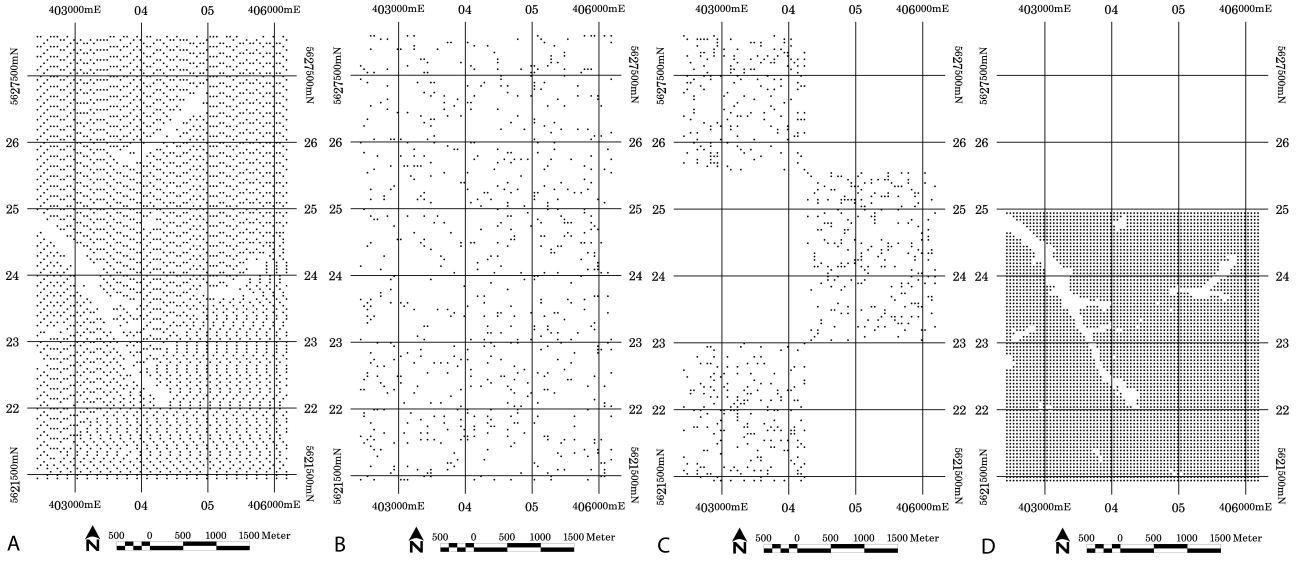


Fig. 5. Different distributions of training data, **(A)** (Uniform): Training set 1 contains 3218 data points, which is one third of all data points available in the study area, **(B)** (Random): Training set 2 comprises 678 data points, **(C)** (Chequerboard): T raining set 3 includes 650 training data points, **(D)** (Localized): Training set 4 contains the largest number of training data, 5751.

classification (refer to Sect. 3.2.2). Considering a coordinate system, each category, e.g. forest, forms one axis.

3.4 Applied accuracy measures

Accuracy assessment was based on validation data sets. As suggested by Foody (2002), for all classifications performed within this study different measures of accuracy, i.e. overall accuracy (OA) and the estimate of Kappa ($\hat{K}$), were calculated and confusion matrices are presented. The calculation of the various accuracy measures was based on equations presented by Congalton and Green (1999). The confusion matrix, also called error matrix, contains information about the real class membership of data points and the class membership assigned to data points by a classification method. Correctly classified points are recorded along the matrix diagonal. Overall accuracy (OA) is the most frequently used accuracy measure and represents the percentage of correctly classified points. $\hat{K}$ takes into account not only the actual agreement, represented by the matrix diagonal, but also the chance agreement, indicated by the row and column totals. $\hat{K}$ values greater than 0.8 represent a strong agreement between the classification and the reference data, whereas values between 0.4 and 0.8 represent moderate agreement and values less than 0.4 stand for poor agreement. The $\hat{K}$ value can be used to determine whether the agreement between a classification and the reference data is significantly greater than 0, meaning better than random labelling, and to test whether two confusion matrices are significantly different. For this purpose Z and Z_p values were calculated (Congalton and Green, 1999, p. 51):

$$Z = \frac{\hat{K}}{\hat{var}(\hat{K})} \tag{11}$$

$$Z_p = \frac{|\hat{K}_1 - \hat{K}_2|}{\sqrt{\hat{var}(\hat{K}_1) + \hat{var}(\hat{K}_2)}} \tag{12}$$

A Z value greater than 1.96 shows that at a confidence level of 95% the classification is significantly better than random. A Z_p value greater than 1.96 means that the two confusion matrices are significantly different, considering a confidence interval of 95%. The equation for the estimated variance of Kappa ($\hat{var}(\hat{K})$) can be found in Congalton and Green (1999, p. 50). Foody (2004) shows that Kappa values can underestimate the accuracy of the classification if the data sets on which the values are estimated are not independent. In the present study, kappa values are measured on very different data sets, as independent as natural data sets can be, and therefore give a good estimation of the classification accuracy. The McNemars test (Bradley, 1968; Agresti, 1996), an alternative proposed by Foody (2004), gave similar results in the present study. Additionally Kappa are standard values in the scientific community. Finally, Liu et al. (2007) show that other estimators, including those proposed by Foody (2004) do not offer alternatives as there is no way to decide which model of chance agreement is correct.

Table 3. SVM classification results for all training site distributions and input variable sets, their applied σ and C as determined by 10-folds CV, and accuracy assessment (see text for details).

Training set	Input domain	C	σ	train error	test error	OA (%)	$\hat{K}$	$v\hat{a}r(\hat{K})$	Z
1	A	64	1	0.1339	0.2193	78	0.72	4.2111e-05	111
	B	16	1	0.1352	0.2214	78	0.72	4.2479e-05	110
	C	256	1	0.0311	0.1726	83	0.78	3.5041e-05	132
2	A	16	1	0.1298	0.2985	70	0.62	3.6197e-05	103
	B	64	4	0.2212	0.2906	71	0.63	3.5628e-05	105
	C	1024	4	0.0796	0.2568	74	0.67	3.3032e-04	117
3	A	64	1	0.0554	0.5255	47	0.32	7.7086e-05	36.8
	C	1024	1	0.0031	0.556	44	0.29	7.7125e-05	33.0
4	A	4	0.25	0.0659	0.6258	37	0.17	1.3761e-04	14.4
	B	64	1	0.0915	0.5812	42	0.29	8.4886e-05	31.5
	C*	16	1	0.0828	0.5931	41	0.23	1.1756e-04	21.1

Table 4. Z_p: Comparison between SVM, MDC and MLC (training set 2).

Input Domain	Method	OA (%)	$\hat{K}$	Pairwise Comparison	Z_p
C*	SVM	73	0.66	SVM vs. MDC	40.56
C*	MDC	42	0.32		
C	SVM	74	0.67	SVM vs. MDC	46.69
C	MDC	39	0.30		
C	MLC	32	0.23		

4 Results

All classifications were significantly better than random, considering a confidence level of 95% (Table 3), and SVM clearly outperformed MDC and MLC (Table 4). As expected, confusion matrices (e.g. Table A3) reveal that most classifications are biased towards classes that where sufficiently represented within the training data sets, meaning class 3, 4 and 5. Despite the fact that all classifications are better than random, the agreement between a classification and the reference data greatly differed depending on the classification method, the distribution of the training sets and the input variables (Table 3, column $\hat{K}$). The best soil type classification was based on a large amount of training data homogeneously distributed over the whole study and on all input variables including coordinates. This classification led to an overall accuracy of 83% and a $\hat{K}$ value of 0.78. The most realistic data set (training set 2), in combination with the input domain C, led to an overall accuracy of 74 % and a $\hat{K}$ value of 0.67.

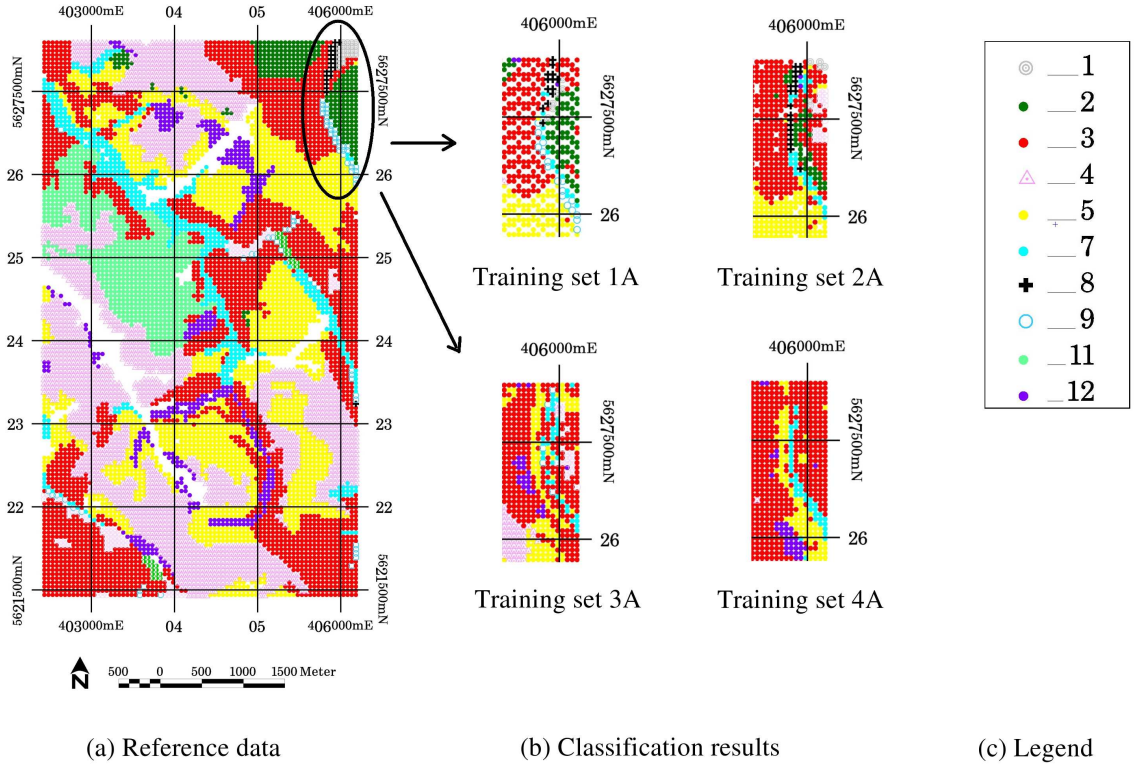
4.1 Classification methods: SVM, MLC and MDC

Table 4 clearly demonstrates that SVM outperformed MLC and MDC. SVM yielded significantly better results than MDC (Z_p>1.96). In addition, the minimum distance classifier led to quite high training errors (input domain C*: e=0.58; input domain C: e=0.62). This indicates a very large class overlap in input space and demonstrates the complexity of the data. This all together supports the choice of SVM as the preferred method.

4.2 Influence of training site distribution

Training sites distributed over the whole study area led to better classification results than training sites that occur rather clustered (training set 3 and 4). The $\hat{K}$ values for classifications performed on training sets 1 and 2 lie between 0.62 and 0.78 and represent moderate agreement. Classifications based on training set 3 and 4 only show poor agreement with the reference data ($\hat{K}$<0.4). They are also more biased towards classes largely represented in the training data set than classifications based on training set 1 and 2. The corresponding confusion matrices are presented in the Appendix (Table A1, A2, A3 and A4).

A higher number of training sites enhanced classification performance considerably (training sets 1 and 2). The $\hat{K}$ value increased from 0.62 to 0.72 (Table 3). However, training set 4 led to the worst classification results (Table 3) despite the large amount of training data. Also a 30-fold CV and a fine-grid search did not enhance the classification. There are mainly two reasons for the bad performance. Training set 4 is the most unbalanced data set, containing classes with more than 1300 data points as well as classes not represented at all (class 1) or represented only by 1, 5 or 20 data points (class 8, 2, 9, 10). However, the main reason is that the altitude of the validation area, ranging from 544 m to 730 m above sea



(a) Reference data (b) Classification results (c) Legend

Fig. 6. Modelling of spatial patterns. Even when the assigned class membership is not correct, the classifier is able to retrieve the correct geometry.

Table 5. Descriptive statistics for the input variable altitude (m above sea level), regarding training set 4 and soil types with more than 200 training data points.

soil types	3	4	5	7	11	12
min	547	601	557	546	613	622
max	816	808	806	791	743	807
mean	691	735	698	608	688	710
median	682	742	693	598	690	703
std	71	45	57	54	31	53
total training data points	1687	1752	1313	228	390	253

level, does not lie within the range of the training data set (721 m to 816 m above sea level). This explains why most data points where assigned to class 3 and class 7, and not to class 4, which is the largest class within the training set (Tables A4, A5 and A6). Soil types 3 and 7 are sufficiently represented within the training data set and are located on lower altitudes (Table 5). Soil type 4, on the other hand, occurs on more elevated places.

All classifications, including classifications based on training set 3 and 4, were able to retrieve certain spatial patterns, as for example the characteristic occurrence of soil types located close to rivers or streams (Fluvisol label:8, Gleysol label:9), in the valley in the north-western part of the study area (Fig. 6). These classes where generally badly represented within the training data sets (Table 3), and especially classifiers trained on training data sets 3 and 4 were not able to assign the correct class memberships to these map units. Instead, these points were allocated to classes sufficiently represented in the training data set, which are partly also located along streams (Gleysol-Planosol label:7). The main point here is that all classifiers assign different class memberships to this region than to the surrounding area. Thus, they reveal the spatial characteristics of soil type distribution in this area.

4.3 Input variables: coordinates

Previous SVM classifications were solely based on coordinates (Pozdnukhov et al., 2002). Within this work coordinates were not considered as the only input variables because there are several advantages to base soil type classification additionally, or solely, on input variables listed above. Firstly, accurate coordinates are not always available. Additionally, using other input variables, it is possible to somehow account for the underlying soil formation processes. Furthermore, the performed classifications show that if the training

Table 6. Z_p: Effect of coordinates on training sets 1 and 2.

training set	input domain	$\hat{K}$	Pairwise Comparison	Z_p
1	A	0.72	A vs. C	7.0338
	C	0.78		
2	A	0.62	A vs. C	6.8103
	C	0.67		

Table 7. Z_p: Effect of coordinates on training sets 3 and 4.

training set	input domain	$\hat{K}$	Pairwise Comparison	Z_p
3	A	0.32	A vs. C	2.6859
	C	0.29		
4	B	0.29	B vs. C*	4.3460
	C*	0.23		

sites are not distributed homogeneously but concentrated in specific parts of the study area, for example due to accessibility problems, coordinates should not be considered as input variables. It is therefore important to test the influence of including the coordinates on the resulting classification.
In case training sites are randomly distributed over the whole study area (training sets 1 and 2), the input of coordinates significantly enhance soil type classification (Table 6).
In case training sites are not homogenously distributed over the whole study area but occur rather clustered in some distinct parts of the study region (training set 3 and 4), classifications based additionally on coordinates were significantly worse than classifications solely based on input variables altitude, slope, aspect, landuse, geology and the distance to rivers (Table 7). The validation data set belonging to training set 4 is located north of the training data set and thus the values of the coordinates of the test data points lie not within the range of the coordinate values of the training data set. This is the same problem as with the input variable altitude presented above. In the test set belonging to training set 3 the coordinate values are within the range of the coordinate values of the training data set. Nevertheless, considering coordinates as input variables deteriorates the classification. The reason is that a classifier based on coordinates is able to account for the spatial distribution of the classes, and in case there are large areas without training data, modelling spatial distribution becomes a disadvantage.

4.4 Input variables: aspect

As expected, for MDC the sine and cosine representation of the aspect yielded significantly better results than the degree repersentation (Z_p=3.59, Table 4). Regarding SVM, no such

Table 8. Z_p: Influence of aspect on training set 1 and 2.

training set	input domain	$\hat{K}$	Pairwise Comparison	Z_p
1	A	0.72	A vs. B	0.3305
	B	0.72		
2	A	0.62	A vs. B	1.0905
	B	0.63		

Table 9. Z_p: Influence of aspect on training set 4.

training set	input domain	$\hat{K}$	Pairwise Comparison	Z_p
4	A	0.17	A vs. B	8.1153
	B	0.29		

clear statement can be formulated. For the classifications based on training sites 1 and 2, the Z_p values are smaller than 1.96, i.e. classifications based on a degree representation of the aspect and classifications based on a sine and cosine representation are not significantly different (Table 8). In this case SVM is able to handle the degree representation of the aspect (refer to Sect. 3.2.2).
Regarding training set 4, however, the classification based on a sine and cosine representation of the aspect (input domain B) is significantly better than the classification based on a degree representation (input domain A) (Table 9). There is no obvious reason why the sine and cosine representation of the aspect significantly enhanced the classification based on training set 4, but did not enhance classification performance for training set 1 and 2. But as Weston et al. declares: "Identifying the reasons why an algorithm works better than another one is difficult".

5 Discussion

The conducted experiments demonstrate that SVMs can be applied for soil type classification and clearly outperform linear classifiers. Training site distribution and the selection of input variables strongly influence the performance of the soil type classification. A realistic training data set and carefully selected input parameters lead to a classification with an overall accuracy of 74% ($\hat{K}$ value of 0.67; training set 2, input domain C). Such results show that the method can be used operationally. We show that the largest set of homogeneously distributed training data points (training set 1) yielded the best results. This might be surprising because one promoted strength of SVM is that it not depends on a good description of the classes and is thus able to derive good results using only a small amount of training data points which

are generally not evenly distributed over the whole study area (Foody and Mathur, 2006). On the other hand soil types vary hugely both locally, within a soil type surface, and regionally, where soils might have a similar type but with different genetic parameters. In such case, it is not only difficult to map soil type boundaries but selective training on these spatial transitions might lead to decreasing performances. For all decision functions the same parameter combination of σ and C was applied in this study (class-insensitive multi-class approach). However, it would be also possible to define C and σ for each decision function separately (class-adaptive approach). Pozdnukhov et al. (2002) state that for their soil texture classification the class-adaptive approach delivered the best results, as it allows somehow to take into account the different spatial variability of classes. Thus, applying a class-adaptive approach here, might enhance classification performance. According to Chen et al. (2005) it is not clear which approach is favourable. In the class-insensitive approach, a uniform parameter combination might not be good for all decision functions, and a class-adaptive approach might lead to overfitting. Working with realistic data sets revealed several problems that need to be discussed here. Within an area of the size of 28 km^2 many soil types, 12 in our case, are present, each varying considerably in their spatial extend. Training data sets are thus also mostly very unbalanced, representing one class with 100 or 1000 times more data points than another one. The second problem is that in some cases training data sets might not contain all classes present within the study area (training set 3 and 4). This is a general problem regarding supervised classifications that assign discrete class memberships to the map units. Assigning probabilities rather than discrete class memberships to map units that are to be classified, would help identifying critical areas. Mantero et al. (2003) explicitly addressed the problem of unkown classes within the study area and suggest a partially supervised classification. Regarding the problem of unbalanced data sets, this study shows that, as expected, classifications are biased towards classes largely represented in the training data set. This again is a problem commonly encountered when working with realistic data sets (e.g. Bhattacharya and Solomatine, 2006). Down-sampling large classes or over-sampling of small classes may help to avoid this problem. However, down-sampling large classes lead to a loss of information, which is critical in case potential SVs are excluded. Over-sampling, on the other hand, may lead to overfitting. Another way to deal with unbalanced data is to penalise misclassifications of samples belonging to small classes more than those of samples belonging to big classes, by assigning higher C values to small and smaller C values to large classes (Osuna et al., 1997, p. 15). A high C value in general would also force the classifier to built an accurate model, but here all classes would be overfitted. Eitrich and Lang (2005) explicitely addressed this problem and presented a way how to tune SVM parameters for large and unbalanced data sets. Regarding the classifications performed within this study, the CV already determined a relatively high C value, and classifiers trained on training set 1 and 2 classified small classes reasonably good. Nevertheless, it would be worthwhile to test whether the methods presented above enhance classification performance for the given data set. Besides the training data sets also the test data sets were unbalanced (Table 2). Thus, overall accuracy mainly evaluates whether large classes are classified correctly. To get around this, accuracy assessment was mainly based on confusion matrices and the $\hat{K}$. $\hat{K}$ was regarded the more appropriate accuracy measure here, as it accounts for chance agreement.

6 Conclusions

The performed classifications show that SVM based on an RBF kernel, the C-soft margin support vector classifier and the One-Against-One multi-class approach can successfully be used to classify soil types. The attributes altitude, slope, aspect, distance to rivers or streams, landuse and geology turned out to be appropriate input variables for soil type classification in the eastern part of the Erzgebirge region. Using coordinates as additional input variables is not always advisable, it strongly depends on the training set distribution. Nevertheless, all these input variables can be derived from remote sensing data or are already widely available, as, for example, the input variable geology. Due to the facility to gather input data, this method is very robust and can be easily used elsewhere.

This study showed, that the best SVM soil type classifications were obtained on the basis of training data, which were more or less homogeneously distributed over the whole study area. Regarding such a distribution of the training data, coordinates enhanced classification performance significantly and are therefore valuable input variables. Thus, given such well distributed training sites, coordinates should generally be considered as input variables. Classifications based on training sites which are not homogeneously distributed over the whole area but occur rather clustered in certain regions of the study area, leaving behind big areas without training sites, only showed poor agreement between reference data and the classification. However, even classifiers trained on such an undesirable training data set were able to detect certain spatial patterns.

Regarding the input variable aspect, it was approved that classifiers based on descriptive statistics (MDC) yield better results using the sine and cosine representation instead of the degree representation. In contrast, SVM should generally be able to handle the degree representation of the aspect, and for two out of three classifications it indeed did not matter whether the degree or the sine and cosine representation was used. However, in one case the the sine and cosine representation produced significantly better results than the degree representation. Thus, it might be worth to try both aspect representations. We propose now to investigate

whether methods concerning the handling of unbalanced data are able to indeed enhance classification results. To account for the fact that the training data sets might not represent all classes within the study region, it would be favourable to assign probabilities to the classified map units, rather than discrete class values and to applied such partially supervised methods as presented by Mantero et al. (2003). Furthermore, additional input variables should be tested, as for example morphology, meaning for instance the position of a map unit on a hill top or in a depression. This attribute might enhance classification performance, because the position on a relief has a strong influence on soil type formation. However, it should be kept in mind that each attribute adds not only additional information but also some noise. Thus, it is favorable to assess the importance of the input variables and to apply feature selection methods. This becomes especially important if a large number of input variables are available.

Appendix A

Confusion Matrices

Confusion matrices calculated for the test data are listed below.

Table A1. Confusion Matrix: Training set 1, input domain A.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	1	0	13	0	0	0	0	1	0	0	0	1
	2	2	101	67	7	6	0	2	1	0	0	0	4
	3	3	32	1631	49	88	4	25	3	9	6	22	15
	4	0	0	36	1370	139	0	4	0	0	0	28	17
	5	1	6	113	178	1094	0	9	0	3	0	15	25
	6	0	1	27	0	0	24	11	0	4	0	0	3
	7	0	4	35	16	46	1	217	0	4	4	12	19
	8	0	0	3	0	0	0	3	9	0	0	0	0
	9	0	1	15	0	1	3	12	0	14	3	0	1
	10	0	0	5	0	0	2	3	0	0	13	0	0
	11	0	0	35	37	19	0	12	0	0	0	416	6
	12	0	4	24	36	56	0	8	0	0	0	1	134

Table A2. Confusion Matrix: Training set 2, input domain A.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	4	0	10	0	2	0	5	0	0	0	0	0
	2	2	37	173	6	6	20	5	2	5	0	0	1
	3	5	18	2234	44	143	43	69	2	8	3	32	28
	4	0	3	70	1659	399	0	3	0	0	0	88	14
	5	0	7	170	309	1380	0	34	0	1	0	42	55
	6	0	0	31	1	1	33	16	0	4	1	0	2
	7	0	1	70	37	70	12	245	2	0	8	13	39
	8	0	0	4	0	0	0	4	11	0	0	0	2
	9	0	0	23	0	7	2	27	5	1	5	0	0
	10	0	2	11	0	0	3	2	0	0	11	0	2
	11	0	0	82	64	49	0	21	0	0	0	523	8
	12	0	1	46	48	91	0	32	0	1	0	2	158

Table A3. Confusion Matrix: Training set 3, input domain A.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	0	0	8	2	9	3	2	0	0	0	0	0
	2	0	0	150	1	26	22	46	0	1	0	0	3
	3	0	11	658	22	163	15	59	0	11	8	54	69
	4	0	12	75	913	331	0	1	0	0	0	13	15
	5	0	1	163	348	546	1	4	0	0	0	12	13
	6	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	6	17	16	3	19	0	2	1	6	9
	8	0	0	11	0	0	0	9	0	2	0	0	0
	9	0	0	25	0	6	0	8	0	2	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0
	11	0	0	57	90	355	0	22	0	0	0	92	7
	12	0	8	51	45	83	0	11	0	2	0	0	29

Table A4. Confusion Matrix: Training set 4, input domain A.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	0	0	24	0	0	0	0	0	0	0	0	0
	2	0	0	220	0	29	0	6	0	0	0	0	17
	3	0	0	924	43	104	8	46	0	0	1	7	10
	4	0	0	315	133	133	0	11	0	0	0	38	6
	5	0	0	610	13	166	1	7	0	0	0	3	41
	6	0	0	12	0	0	0	9	0	0	7	0	0
	7	0	0	216	0	9	0	86	0	0	2	1	0
	8	0	0	15	0	0	0	7	0	0	0	0	0
	9	0	0	12	0	10	2	26	0	0	0	0	0
	10	0	0	8	0	0	0	6	0	0	1	0	0
	11	0	0	121	57	65	0	22	0	0	0	132	3
	12	0	0	129	0	13	0	4	0	0	0	2	9

Table A5. Confusion Matrix: training set 4, input domain B.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	0	0	9	0	0	2	13	0	0	0	0	0
	2	0	0	222	5	12	0	15	0	1	0	0	17
	3	0	0	715	73	119	25	169	0	2	2	24	14
	4	0	0	53	284	112	0	53	0	0	0	85	49
	5	0	1	132	118	212	3	145	0	0	0	34	196
	6	0	0	2	0	0	0	12	0	0	14	0	0
	7	0	0	56	2	20	0	207	0	1	7	10	11
	8	0	0	1	0	0	0	20	0	1	0	0	0
	9	0	0	6	0	7	4	32	0	0	1	0	0
	10	0	0	8	0	0	0	4	0	0	3	0	0
	11	0	0	36	52	76	0	51	0	0	0	183	2
	12	0	0	54	13	19	3	31	0	0	0	17	20

Table A6. Confusion Matrix: training set 4, input domain C.

		Classification											
		1	2	3	4	5	6	7	8	9	10	11	12
Ground reference	1	0	0	22	0	1	0	1	0	0	0	0	0
	2	0	0	249	0	15	0	8	0	0	0	0	0
	3	0	0	899	0	130	11	46	0	0	7	44	6
	4	0	0	384	64	63	0	17	0	0	0	76	32
	5	0	0	428	3	275	0	47	0	0	0	29	59
	6	0	0	3	0	1	0	5	0	0	19	0	0
	7	0	0	181	0	10	0	101	0	0	10	10	2
	8	0	0	1	0	0	0	21	0	0	0	0	0
	9	0	0	7	0	11	4	27	0	0	1	0	0
	10	0	0	6	0	0	0	6	0	0	3	0	0
	11	0	0	131	16	26	0	11	0	0	0	207	9
	12	0	0	58	0	33	2	28	0	0	0	7	29

Acknowledgements. We thank H. Gaonach and two anonymous referees for their careful and constructive reviews. The data were provided by EMTAL project group. We used the free software SPIDER to perform SVM classifications. The SPIDER toolbox is an object orientated environment for machine learning in Matlab (http://www.kyb.mpg.de/bs/people/spider/main.html). We thank H. Pöhler who kindly answered all questions about the data, and G. H. Baklir who provided support regarding the use of the SPIDER toolbox. We also thank H. Schaeben for his support regarding SVM theory.

Edited by: H. Gaonach
Reviewed by: two anonymous referees

References

Agresti, A.: An Introduction to Categorical Data Analysis, Wiley Press, p. 312, 1996

Bhattacharya, B. and Solomatine, D. P.: Machine learning in soil classification, Neural Networks, 19, 186–195, 2006.

Bradley, J. V.: Distribution-Free Statistical Tests, Prentice-Hall Press, p. 388, 1968.

Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition, Data Min. Knowl. Disk., 2, 121–167, 1998.

Chen, P.-H., Lin, C.-J., and Schölkopf, B.: A Tutorial on nu-Support Vector Machines, Appl. Stoch. Model. Bus., 21, 111–136, 2005.

Congalton, R. G. and Green, K.: Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Lewis Publishers, p. 160 , 1999.

Cortes, C. and Vapnik, V. N.: Support vector networks, Machine Learning, 20, 273–297, 1995.

Eitrich, T. and Lang, B.: Parallel Tuning of Support Vector Machine Learning Parameters for Large and Unbalanced Data Sets, Lecture Notes in Computer Science 3695, Springer-Verlag 2005, Konstanz (Germany, 2005), 1st International Symposium on Computational Life Sciences (CompLife'05), p. 253–264, 2005.

Foody, G.: Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy, Photogramm Eng. Rem. S., 70, 627–633, 2004.

Foody, G. M.: Status of land cover classification accuracy assessment, Remote Sense Environ., 80, 185–201, 2002.

Foody, G. M. and Mathur, A.: A Relative Evaluation of Multiclass Image Classification by Support Vector Machines, IEEE T. Geosci. Remote, 42, 1335–1343, 2004.

Foody, G. M. and Mathur, A.: The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM, Remote Sens. Environ., 103, 179–189, 2006.

Hsu, C. W. and Lin, C. J.: A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, 13, 415–425, 2002.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J.: A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, accessed: 14 August 2007, 2007.

Huang, C., Davis, L. S., and Townshend, J. R. G.: An assessment of support vector machines for land cover classification, Int. J. Remote Sens., 23, 725–749, 2002.

Kanevski, M., Pozdnoukhov, A., Timonin, V., and Maignan, M.: Mapping of environmental data using kernel based methods, Research Report, p.18, 2007.

Liu, C., Fraizier, P., and Kumar, L.: Comparative assessment of the measures of thematic classification accuracy, Remote Sens. Environ., 107, 606–616, 2007.

Mantero, P., Moser, G., and Serpico, S.: Partially supervised classification of remote sensing images using SVM-based probability density estimation, Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on, pp. 327–336, 2003.

Osuna, E., Freund, R., and Girosi, F.: AI Memo 1602, Massachusetts Institute of Technology, 1997.

Pal, M. and Mather, P. M.: Support vector machines for classification in remote sensing, Int. J. Remote Sens., 26, 1007–1011, 2005.

Pozdnukhov, A., Timonin, V., Kanevski, M., Savelieva, E., and Chernov, S.: Classification of Environmental Data with Kernel Based Algorithms, Preprint IBRAE, moscow: Nuclear Safety Institute RAS, p. 22, 2002.

Schölkopf, B. and Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, p. 664, 2002.

Shawe-Taylor, J. and Cristianini, N.: An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, p. 189, 2000.

Vapnik, V. N.: The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science), Springer Verlag, second edn., p. 314, 2000.

Weston, J. and Watkins, C.: Multi-class Support Vector Machines, technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK, p. 10, 1998.

Weston, J., Gretton, A., and Elisseeff, A.: SVM Practical (How to get good results without cheating), http://www.kyb.mpg.de/bs/people/spider/main.html, accessed: 17 January 2007.